ERLHS: A Hamiltonian Framework for Coherence-Preserving Machine Intelligence

Sylvain Cormier
Paraxiom Research
sylvain@paraxiom.org

December 2025

Abstract

Current large language models (LLMs) operate without geometric or physical constraints on their latent dynamics. As a consequence, arbitrary textual perturbations—including prompt injection attacks—can drive their internal states into regions never encountered during training, resulting in incoherence, contradiction, and a lack of robust continual learning.

We introduce ERLHS (Externally-Regularized Latent Hamiltonian Systems), a framework in which latent representations evolve on a smooth manifold equipped with a Hamiltonian coherence functional. Valid transitions are those that preserve or reduce this functional, providing a physically-motivated invariant that constrains updates. We formalize prompt injection as an *off-manifold perturbation problem* and show why conventional LLMs cannot defend against it. ERLHS ensures coherence preservation, bounded adversarial influence, and safe continual learning. A distributed proof-of-coherence mechanism provides transition integrity across nodes without relying on game-theoretic assumptions.

1 Background and Problem Formulation

1.1 Latent Spaces Without Manifold Structure

Modern deep architectures operate in $\mathbb{R}^n$ with no enforced manifold structure. Let $h_t \in \mathbb{R}^n$ denote a hidden state updated by

$$h_{t+1} = f_\theta(h_t, x_t).$$

There is no requirement that h_{t+1} remain in a geometrically meaningful subspace. This lack of structure allows arbitrary inputs to redirect trajectories.

1.2 Hamiltonian Systems

A Hamiltonian system is defined on a smooth manifold M equipped with a symplectic form ω. Given a Hamiltonian function $H : M \to \mathbb{R}$, the dynamics follow

$$\dot{q} = \frac{\partial H}{\partial p}, \qquad \dot{p} = -\frac{\partial H}{\partial q}.$$

In intrinsic form, the Hamiltonian vector field X_H satisfies

$$\iota_{X_H}\omega = dH.$$

This structure enforces invariants, bounded energy, and coherent flow [**?**].

1.3 Coherence Functional

We define a *coherence functional* $H : M \to \mathbb{R}$ that penalizes off-manifold drift. Intuitively, H measures deviation from learned relationships among latent variables. Coherent reasoning corresponds to trajectories of non-increasing H.

1.4 Prompt Injection as Off-Manifold Perturbation

In LLMs, a perturbation $x_t \mapsto x'_t$ induces

$$h'_{t+1} = f_\theta(h_t, x'_t),$$

with no constraint ensuring $h'_{t+1} \in M$.

Thus prompt injection is the problem of forcing the internal trajectory outside any region where model behavior is predictable or trained [?]. Existing defenses (RLHF, filters) intervene only in output space and do not constrain the latent geometry [?].

2 The ERLHS Framework

2.1 Definition

An ERLHS agent is a tuple

$$(M, \omega, H, T, \mathcal{C}),$$

where:

- M is a smooth latent manifold,
- ω is a symplectic structure,
- $H : M \to \mathbb{R}$ is a coherence functional,
- T is a transition operator on M,
- $\mathcal{C}$ is a coherence verifier.

2.2 Transition Constraint

A transition $z_t \mapsto z_{t+1}$ is admissible iff

$$H(z_{t+1}) \leq H(z_t) + \varepsilon,$$

for a small tolerance ε. This enforces coherence-preserving flow.

2.3 Coherence Verification

The verifier $\mathcal{C}$ checks that T satisfies the Hamiltonian constraint. Invalid transitions are rejected before they can propagate to downstream reasoning.

3 Limitations of Existing Models

3.1 No Symplectic Geometry

Standard architectures do not enforce ω-structure. Their dynamics lack invariants and cannot resist adversarial redirection.

3.2 No Hamiltonian Invariant

Without conserved quantities, latent evolution is unconstrained.

3.3 Off-Manifold Vulnerability

Prompt injection is possible precisely because LLMs have no mechanism to ensure that h_{t+1} lies on any learned manifold M. Formally:

$$f_\theta(h_t, x'_t) \notin M \quad \Rightarrow \quad \text{unbounded drift.}$$

3.4 RLHF Insufficiency

RLHF shapes output distributions but does not impose geometric structure on the latent space or constrain transition dynamics.

4 Robustness Properties of ERLHS

4.1 Bounded Adversarial Influence

If H is Lipschitz with constant L_H, then

$$\|z_{t+1} - z_t\| \leq L_H^{-1} |H(z_{t+1}) - H(z_t)|.$$

Adversarial perturbations cannot induce large hidden-state deviations.

4.2 Coherence Preservation

Hamiltonian dynamics ensure smooth, reversible, and stable flow across M. Reasoning cannot "jump" into incoherent configurations.

4.3 Safe Continual Learning

Bounding ΔH prevents catastrophic forgetting and uncontrolled parameter drift [?]. Updates respect learned geometric structure.

5 Distributed Proof of Coherence

5.1 Consensus on Invariants

Nodes verify that proposed transitions satisfy the coherence constraint. Consensus ensures global agreement on allowed latent evolution.

5.2 Quantum Authentication

QKD or quantum-resistant signatures authenticate transitions, ensuring that no adversary can forge coherent trajectories [?].

5.3 Physics-Based Validation

The ledger does not validate semantic content, only that coherence invariants are preserved. This is strictly simpler and more robust than game-theoretic consensus.

6 Implementation Notes

ERLHS can be layered atop existing architectures by projecting latent vectors onto M, enforcing Hamiltonian updates, and inserting coherence checks between modular components. Hamiltonian neural networks provide a foundation for learning structure-preserving dynamics [**?**].

6.1 Scope

This work defines a governance and stability formalism for latent dynamics in machine learning systems. It provides theoretical foundations for coherence-preserving architectures rather than empirical benchmarks. Experimental evaluation comparing ERLHS-constrained models against baseline LLMs on adversarial robustness and continual learning tasks is the subject of ongoing and future work.

7 Conclusion

LLMs fail under prompt injection because they lack geometric constraints and Hamiltonian invariants. ERLHS introduces manifold structure, coherence functionals, and physically-motivated transition constraints that eliminate this vulnerability and enable stable continual learning.

References

[1] V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 2nd edition, 1989.

[2] D. J. Bernstein and T. Lange. Post-quantum cryptography. *Nature*, 549(7671):188–194, 2017.

[3] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. In *NeurIPS*, 2019.

[4] J. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

[5] L. Ouyang et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[6] F. Perez and I. Ribas. Ignore this title and hack a prompt injection primer. *arXiv preprint arXiv:2211.09527*, 2022.